IMPROVEMENT AND MEASUREMENT OF NEURAL STYLE TRANSFER

BY

MAO-CHUANG YEH

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Adviser:

David A. Forsyth

ABSTRACT

Style transfer methods produce a transferred image which is a rendering of a content image in the manner of a style image. There is a rich literature of variant methods. We seek to understand how to improve style transfer: in particular, there is some evidence that cross-layer losses are helpful, and some evidence that optimization problems might present difficulties. To do so requires quantitative evaluation procedures, but current evaluation is qualitative, mostly involving user studies. We describe a novel quantitative evaluation procedure. Our procedure relies on two statistics: the Effectiveness (E) statistic measures the extent that a given style has been transferred to the target, and the Coherence (C) statistic measures the extent to which the original image's content is preserved. Our statistics are calibrated to human preference: targets with larger values of E (resp C) will reliably be preferred by human subjects in comparisons of style (resp. content).

We use these statistics to investigate relative performance of a number of recent style transfer methods, revealing a number of intriguing properties. Our experiments pool multiple style transfers from many different styles to many different content images using many different style weights, allowing us to make general statements about what influences style transfer. Admissible methods lie on a Pareto frontier (i.e. improving E reduces C, or vice versa). Three methods are admissible: Universal style transfer produces very good C but weak E; modifying the optimization used for Gatys' loss produces a method with strong E and strong C; and a modified cross-layer method has slightly better E at strong cost in C. While the histogram loss improves the E statistics of Gatys' method, it does not make the method admissible. Surprisingly, style weights have relatively little effect, and most variability in transfer is explained by the style itself (meaning experimenters can be misguided by selecting styles).

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Neural style transfer methods apply the *style* from one example image to the *content* of another; for instance, one might render a camera image (the content) as a watercolor painting (the style). This is done by constructing an image so that the statistics of some network layers match those of the style while other layers directly match those of the content image. The standard is due to Gatys [1]. After Gatys, there are many algorithm for improving neural style transfer. Novak and Nikulin [2] first mentioned cross-layer statistics within a wide range of variant style transfer methods though their purpose and focus is not exploring and understanding of cross-layer statistics.

At this stage, to study the differnece between Corss-layer statistics and Within-layer statistics is an important topic not only for style transfer but also for helping understanding the neural network. In this thesis, we study the properties of Cross-layer statistics i.e. Cross-layer gram matrix. We theoretially study its loss in Chapter 2 and Apppendix D, and provide qualitative comparison to support the arguments in Chapter 2. On the other hand, we seek to identify factors that lead to strong style transfers. To do so, in Chapter 3 we construct a comprehensive quantitative evaluation procedure for style transfer methods. We evaluate style transfers on two criteria. **Effectiveness** (E) measures whether transferred images have the desired style, using divergence between Convolutional Neural Network (CNN) feature layer distributions of the synthesized image and original image. **Coherence** (C) measures whether the synthesized images respect the underlying decomposition of the content image into objects, using established procedures together with the Berkeley segmentation dataset BSDS500 [3]. Both our E and C measures are calibrated by user studies.

Contributions: We are the first proving cross-layer gram matrix works better than within-layer gram matrix for style transfer. Experimentaly, we have both qualitaive reuslt and quantitaitve reuslt. We present E and C measures of style transferred images (see Fig. 1.1). Our measures are highly effective at predicting user preferences. We use our measures to compare several style transfer methods quantitatively. Our study suggests that controlling cross-layer loss is helpful, particularly if one uses the cross-layer covariance matrix (rather than Gram matrix). Our study suggests that, despite the analysis of Risser et al. [4], the main problem with Gatys' method is optimization rather than symmetry; modifying the optimization leads to an extremely strong method. Gatys' method is unstable with high style weights, and we construct explicit models of the symmetry groups for Gatys' style loss and the cross-layer style loss (improving over Risser *et al.* , who could not construct the groups), which may explain this effect. Our study suggests that, even for the best methods

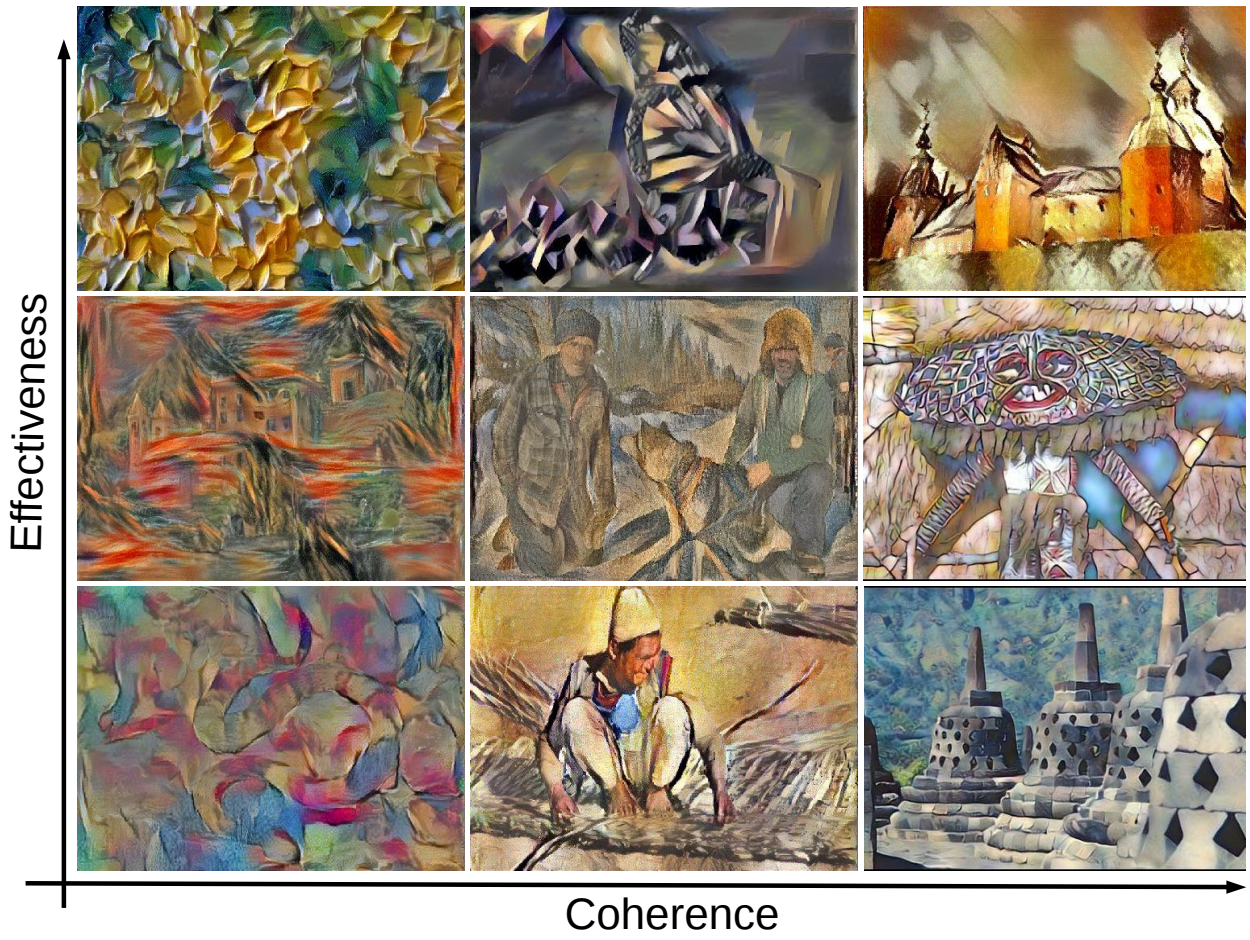


Figure 1.1: *A grid of transfers visualizing Effectiveness-Coherence space. We break Effectiveness (resp. Coherence) into three quantiles, then show an image selected from each grid box. Coherence increases from left to right, and Effectiveness increases from bottom to top, as in the graphs.*

we investigated, the effect of choice of style image is strong, meaning that it is dangerous for experimenters to select style images when reporting results.

1.1 RELATED WORK

Style transfer: bilinear models [5] , non-parametric methods [6], image analogies [7] and adjusting filter statistics [8, 9] are capable of image style transfer and yield texture synthesis. Gatys *et al.* demonstrated that producing neural network layers with particular summary statistics (i.e. Gram matrices) yielded effective texture synthesis [10]. Gatys *et al.* achieved style transfer by searching for an image that satisfies both style texture summary statistics and content constraints [1]. This work has been much elaborated [11, 12, 13, 14,

15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. Novak and Nikulin noticed that cross-layer Gram matrices reliably produce improvement on style transfer ([2]). However, their work was an exploration of variants of style transfer rather than a thorough study to gain insights on style summary statistics; since then, the method has been ignored in the literature.

Style transfer evaluation: style transfer methods are currently evaluated mostly by visual inspection on a small set of different styles and content image pairs. To our knowledge, there are no quantitative protocols to evaluate the competence of style transfer apart from user studies [18] (who also investigate edge coherence between content and stylized images).

Gram matrices symmetry in a style transfer loss function occur when there is a transformation available that changes the style transferred image without changing the value of the loss function. Risser *et al.* note instability in Gatys' method; symptoms are: poor and good style transfers of the same style to the same content with about the same loss value [4]. They supply evidence that this behavior can be controlled by adding a histogram loss, which breaks the symmetry. They do not write out the symmetry group as too complicated ([4], p 4-6). Gupta *et al.* [26] link instability in Gaty's method to the size of the trace of the Gram matrix.

1.2 METHOD AND NOTATION

We review the original work of Gatys *et al.* [1] in detail to introduce notation. Gatys finds an image where early layers of a CNN representation match the lower layers of the style image and higher layers match the higher layers of a content image. Write I_s (resp. I_c, I_n) for the style (resp. content, new) image, and α for some parameters balancing style and content losses (L_s and L_c respectively). Occasionally, we will write $I_n^m(I_c, I_s)$ for the image resulting from style transfer using method m applied to the arguments. We obtain I_n by finding

$$\underset{I_n}{\operatorname{argmin}} L_c(I_n, I_c) + \alpha L_s(I_n, I_s)$$

Losses are computed on a network representation, with L convolutional layers, where the l'th layer produces a feature map f^l of size $H^l \times W^l \times C^l$ (resp. height, width, and channel number). We partition the layers into three groups (style, content and target). Then we reindex the spatial variables (height and width) and write $f^l_{k,p}$ for the response of the k'th channel at the p'th location in the l'th convolutional layer. The content loss L_c is

$$L_c(I_n, I_c) = \frac{1}{2} \sum_c \sum_{k,p} \left\| f^c_{k,p}(I_n) - f^c_{k,p}(I_c) \right\|^2 \tag{1.1}$$

(where c ranges over content layers). The *within-layer gram matrix* for the l'th layer is

$$G^l_{ij}(I) = \sum_p \left[f^l_{i,p}(I) \right] \left[f^l_{j,p}(I) \right]^T . \tag{1.2}$$

Write w_l for the weight applied to the l'th layer. Then

$$L^l_s(I_n, I_s) = \frac{1}{4N^{l2}M^{l2}} \sum_s w_l \sum_{i,j} \left\| G^s_{ij}(I_n) - G^s_{ij}(I_s) \right\|^2 \tag{1.3}$$

where s ranges over style layers. Gatys *et al.* use Relu1_1, Relu2_1, Relu3_1, Relu4_1, and Relu5_1 as style layers, and layer Relu4_2 for the content loss, and search for I_n using L-BFGS [27]. From now on, we write R51 for Relu5_1, etc.

CHAPTER 2: CROSS-LAYER STYLE LOSS AND QUALITATIVE COMPARISON

Novak and Nikulin noticed that across-layer gram matrices reliably produce improvement on style transfer. ([2]). However, their work was an exploration of variants of style transfer rather than a thorough study to gain insights on style summary statistics. There are reasons cross-layer terms produce improvements. In some styles, very long scale patterns are formed out of small components. For instance, in Figure 2.2, small white spots are organized into long curves. Within-layer gram matrices are not well adapted to represent this phenomenon, as Figure 2.2 shows. Generally, such hard styles occur where effects at short spatial scales are organized into longer scale structures. Such hard styles are strongly associated with physical materials. In this chapter, we show that comparing cross-layer gram matrices – which encode co-occurrences between (say) small and medium scale patterns — produces qualitive improvements in style transfer for such styles. Furthermore, controlling cross-layer gram matrices also effectively controls pattern frequencies.

We consider a style loss that takes into account between layer statistics. The **cross-layer, additive (XL)** loss is obtained as follows. Consider layer l and m, both style layers, with decreasing spatial resolution.

Write $\uparrow f^m$ for an upsampling of f^m to $H^l \times W^l \times K^m$, and consider

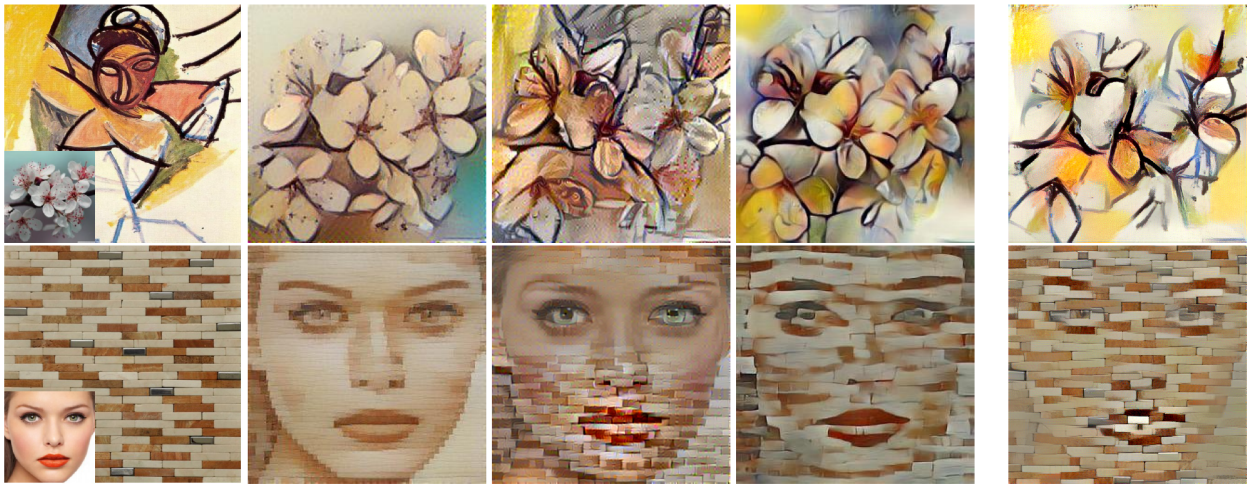
$$G_{ij}^{l,m}(I) = \sum_p \left[f_{i,p}^l(I)\right] \left[\uparrow f_{j,p}^m(I)\right]^T. \tag{2.1}$$

as the cross-layer gram matrix, We can form a style loss

$$L_s(I, I_s) = \sum_{(l,m)\in\mathcal{L}} w^l \sum_{ij} \left\| G_{ij}^{l,m}(I) - G_{ij}^{l,m}(I_s) \right\|^2 \tag{2.2}$$

(where $\mathcal{L}$ is a set of pairs of style layers). We can substitute this loss into the original style loss, and minimize as before. All results here used a *pairwise descending* strategy, where one constrains each layer and its successor (i.e. (R51, R41); (R41, R31); etc). Alternatives include an *all distinct pairs* strategy, where one constrains all pairs of distinct layers. Carefully controlling weights for each layer's style loss is not necessary in cross-layer gram matrix scenario.

Style layer pairs: In principle, any set of pairs can be used. We have investigated a *pairwise descending* strategy, where one constrains each layer and its successor (i.e. (R51, R41); (R41, R31); etc) and an *all distinct pairs* strategy, where one constrains all pairs of



(a) Styles and contents (b) Arbitrary Style [11] (c) Gatys [8] (d) Universal style [14] (e) ours

Figure 2.1: The first column are style and content images; images in column 2,3,4 are results from [16], [1], and [17], which are reported in Y. Li et al. [17]. Our results are at the last column, which use cross-layer gram matrices as style losses and are optimized on multiplicative loss between content and style.

distinct layers.

Pattern management across scales: Controlling within-layer gram matrices by proper weighting ensures that the statistics of patterns at a particular scale are "appropriate". However, we speculate – and our experimental results seem to confirm – that one can get these statistics right without having desirable weighting relations across scales. Inter-layer gram matrices require that phenomena at one scale are correlated to those at the next scale appropriately. In other words, carefully controlling weights for each layer's style loss is not necessary in cross-layer gram matrix scenario.

Number of constraints: Cross-layer gram matrices control considerably fewer parameters than within layer gram matrices. For a pairwise descending strategy, we have four cross-layer gram matrices, leading to control of $64\times128+128\times256+256\times512+512\times512 = 434176$ parameters; compare within layer gram matrices, which control $64^2+128^2+256^2+2\times512^2 = 610304$ parameters. It may seem that there is less constraint on style. Experiment suggests our method produces visible improved results, meaning that many of the parameters controlled by within-layer gram matrices have no particular effect on the outcome.

Figure 2, 2.2, 2.3, provide quantative comparison of how strong of cross-layer in preserving the style in scales.

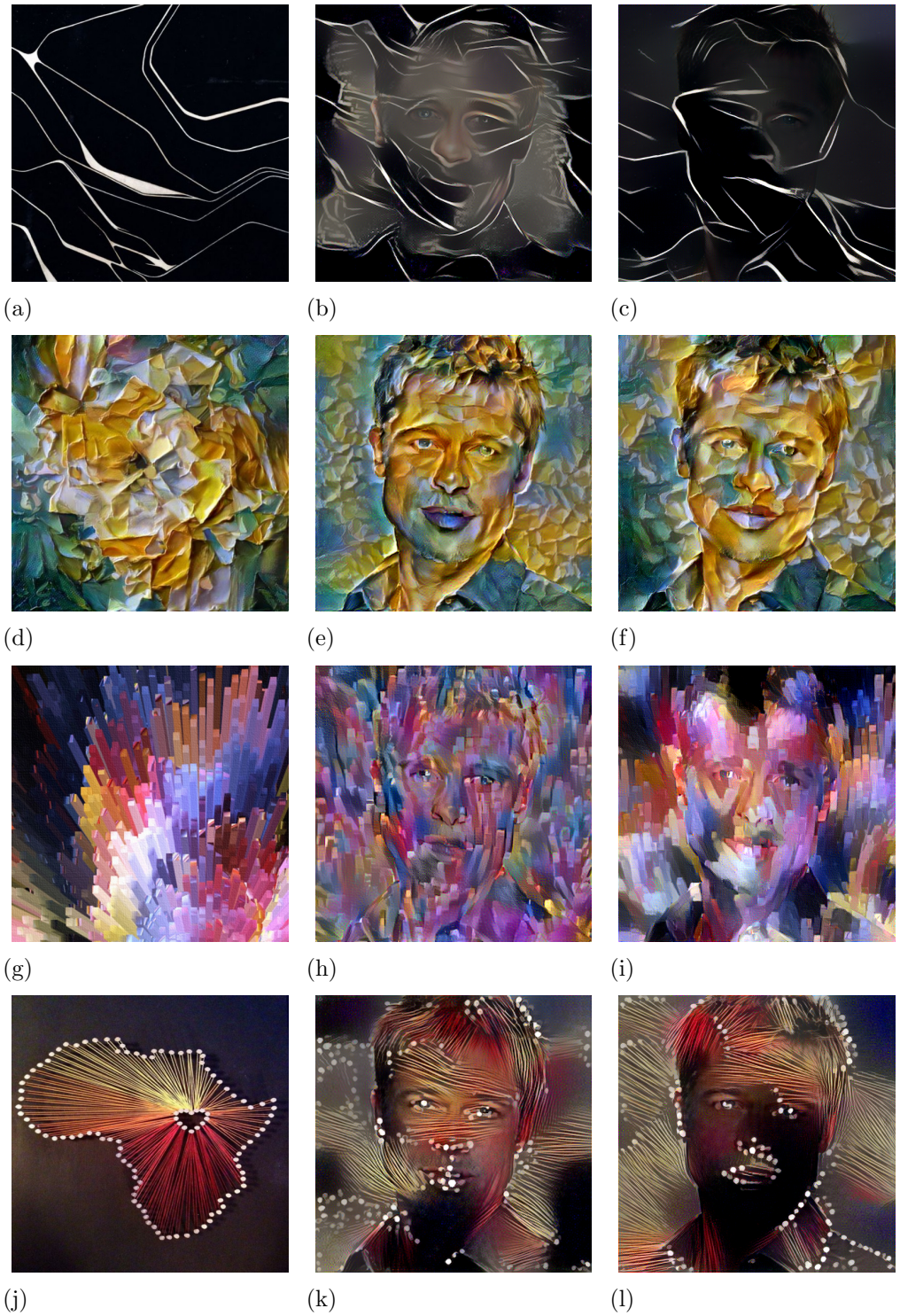


Figure 2.2: **Left:** *styles to transfer;* **center:** *results using within-layer loss;* **right** *results using cross-layer loss. There are visible advantages to using the cross-layer loss. Note how cross-layer preserves large black areas (top row); creates an improved appearance of relief for the acrylic strokes (second row); preserves the overall structure of the rods (third row); and ensures each string has a dot on each end (fourth row).*

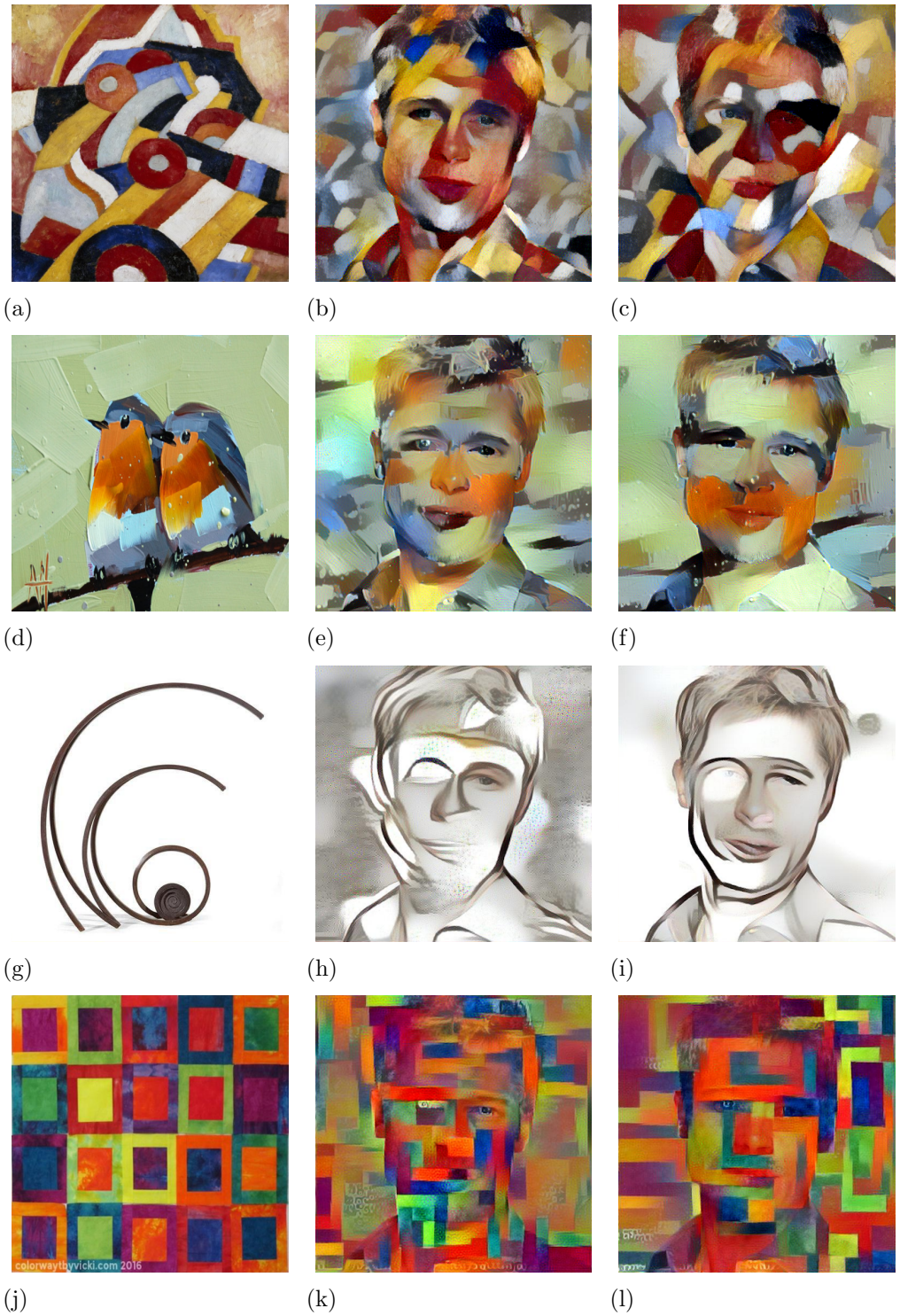


Figure 2.3: **Left:** *styles to transfer;* **center:** *results using within-layer loss;* **right:** *results using cross-layer loss. There are visible advantages to using the cross-layer loss. Note how cross-layer preserves the shape of the abstract color blocks (top row); avoids smearing large paint strokes (second row); preserves the overall structure of the curves as much as possible (third row); and produces color blocks with thin boundaries (fourth row).*

CHAPTER 3: QUANTITATIVE EVALUATION

3.1 BASE STATISTICS FOR QUANTITATIVE EVALUATION

Ideally, a style transfer method should meet two basic tests: (1) the method produces images in the desired style – **E**; (2) the resulting images respect the underlying decomposition of the content image into objects – **C**. While final judgment should belong to the artist, we construct numerical proxies that can be used to disqualify inadmissible methods. Note that analysis on both E and C properties is essential, due to the trade-off characteristics in between: e.g. excellent results on coherence can be obtained with no transfered style at all, and vice versa. In this section, we introduce the base statistics for each measure, which can then be applied to build predictors of human preferences in the user study (see Sec 3.2).

Base E statistics. We consider a style to be applied to a content image. Recent approach [28] reveals that the distribution of features within lower feature layers of a CNN representation is an effective proxy to capture styles. We expect the ideal transfer on each individual image has small biases in the distribution of feature layers to account for the content; and on the scale of a sets of images, the distribution of features should reflect the style distribution itself. We consider a stronger E response of a style transfer method for a particular image as: the distribution of feature layer values produced by the transferred image matches the corresponding distribution of the style image. In notation, write $\mathbf{f}_p^l(I)$ for the vector of responses of all channels at the p'th location in the l'th CNN layer for image I. Given the i'th content image, the j'th style, and some method m. For the transferred image I_n, the distribution P_n^m of $\mathbf{f}^l(I_n^m(I_c^i, I_s^j))$ should be similar to the distribution P_s of $\mathbf{f}^l(I_s^j)$, with possible smoothing effects to meet content demands.

Measuring whether two datasets come from the same, unknown, distribution in high dimensions remains challenging: we do not expect the distributions to be exactly the same; instead, we want to identify obvious (and so suspicious) large differences. We thus compute a set of principal components of layer responses based on within layer covariance matrix averaged over 200 content images through VGG net. We then choose the principle component dimensions of layers R11, R21, R31, R41, R51 respectively with 18, 100, 128, 280, 256. The mean and the covariant matrix of each of the style layer can then be computed. We project the transfer layers' feature onto these principle components to obtain the KL divergences between the style and the transfered one. We obtain in total 5 KL divergences statistics from the 5 layers. In notation, E_i denotes the negative log KL divergence between the i'th layer of the transferred image and the i'th layer of the style image. The dimensionality for

principle component is decided based on two concerns: (1) transfered image visually has better style quality should have lower KL divergence (2) keeping the dimensionality as small as possible to reduce the numerical error of calculating KL divergence.

Base C statistic measures the extent to which it preserves object boundaries in the content image. A style transfer method that eliminates object contours would make it hard for humans to interpret the content. Therefore, we model the coherence of transfered images as the ability to infer cues of object contours. We analyze C statistics on the Berkeley segmentation dataset BSDS500 [3]. We evaluate on the test split only. We apply each style transfer method on the test image to obtain the synthesized one. We then use an existing contour detector by Arbelaez et al. [3]) to capture boundaries and compare it with ground truth. Then for each method, we obtain the probability of boundary (Pb) precision-recall curve for each transfered image. We could then compute the area under curve (AUC) metrics for all methods. A higher AUC suggests better boundary preservation.

3.2 CALIBRATED MEASURES FROM BASE STATISTICS

Our base statistics offer reasonable measurement, but cannot be used directly to rank style transfer methods. One has to calibrate these measures with actual human preference before using them to search strong style transfer methods. Therefore, we involve two user studies (E-test for style and C-test for content, Fig. 3.1) to help calibrate the base measures. In both of the study, users are presented with two transfered images using two different methods, while the content and the style are the same. In the style study, users are asked to choose the transfered image that better captures the style. The transfered images are not pre-selected, meaning that some pairs might have very large difference in E statistics and others might be small. In the content study, users are asked to choose the image that better captures the content. The provided image pairs are pre-qualified for the content study, as they are chosen to have relatively high E statistics (details below). The pre-qualification is achieved with the manual selection about the content faithfulness (a style transfer is *known* not to have worked). Pilot studies provided evidence that human preferences could be accurately predicted using our base effectiveness statistics.

3.3 CALIBRATION WITH USER STUDIES

Calibration is important. e.g. base C statistics may not be particularly reliable: heavily textured styles, which can be distinguished by humans, may confuse the quantitative contour

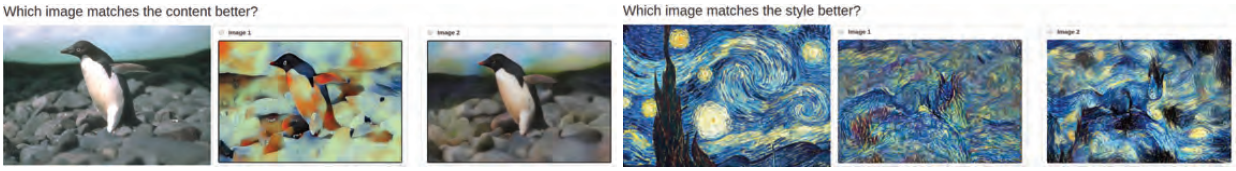


Figure 3.1: *On the* **left**, *a typical screen from the* content *user study; a user must select which target has content most like the given content image. On the* **right**, *a typical screen from the* style *user study; a user must select which target has style most like the given style image. In the content user study, transfers are pre-qualified to have reasonably good E values.*

evaluations. This is because the contour detector was not built with very aggressive texture fields in mind (compare typical style transfer images with the "natural" textures used to build BSDS500). Moreover, we might have texture fields of one object that are strongly coherent visually within each regions but obviously different between regions, while the contour detector has great difficulty in identifying the boundaries. For calibration, we use logistic regression to construct per-image measures that directly predict human preferences from the base statistics.

Calibration: We compare transferred images by comparing the scores they produce from their E and C values. The difference of scores between two transfered images (image 1 and 2) will predict the probability that one is preferred by the human viewer over the other. We obtain such scores using logistic regression. The scores are thus *calibrated* if the predictions of preference are accurate. e.g. if image 1 has score s_1, then the probability that image 1 will be preferred by a user is predicted by $e^{s_1}/(e^{s_1} + e^{s_2})$. We seek one such score for effectiveness (which should predict the results of the style study) and another for coherence (which should predict the results of the content user study).

Controls: We have two controls that are important in calibrating scores. In the first, the resized style image is reported as a transfered image. In the second, the content image is reported as a transfered image.

Scores and models: For each image, we have a random variable y says if this image is referred by human from an transferred image pair, we also have a vector of features $\mathbf{x}$ chosen from some combination of the base C statistic and the 5 base E statistics. Given a pair of images ($\mathbf{x}_1$ for image 1, etc.), we can fit the logistic regression model

$$\frac{\log P(y_1 = 1|\theta, \mathbf{x}_1, \mathbf{x}_2)}{\log P(y_1 = 0|\theta, \mathbf{x}_1, \mathbf{x}_2)} = \theta^T(\mathbf{x}_1 - \mathbf{x}_2) \tag{3.1}$$

Model	Admissible	Cross-validated accuracy
1	yes	.856 (3e-3)
2	yes	.867 (2e-3)
3	yes	.873 (3e-3)
4	no; (b)	.871 (3e-3)
5	no; (b)	.873 (2e-3)

Table 3.1: *Cross validated accuracy for our E-model predictions of human preference in the style experiment, using models described in the text (parens give standard error of cross-validated accuracy).*

Model	Admissible	Cross-validated accuracy
C	yes	.692 (8e-3)
1	yes	.694 (8e-3)
2	no; (b)	.710 (7e-3)
3	no; (b)	.756 (7e-3)
4	no; (b)	.759 (7e-3)
5	no; (b)	.767 (7e-3)

Table 3.2: *Cross validated accuracy for our C-model predictions of human preference in the content experiment, using models described in the text (parens give standard error of cross-validated accuracy).*

which yields a per-image score $s = \theta^T \mathbf{x}$. The choice of the admissible model is important: (a) the model should predict human preferences accurately; (b) the model should be as small as reasonably possible; (c) improvements in any E base statistic should never make an image less preferable in a style test; (d) improvements in the C base statistic should never make an image less preferable in a content test; (e) the model should very strongly prefer content controls to style controls in content tests (and vice versa in style tests).

E statistic: We investigated five E-models, where the r'th uses $\{E_1 \dots E_r\}$. Table 3.1 shows the cross-validated accuracy of the models and whether they are admissible or not. We use the admissible model with $r = 3$, which has highest cross-validated accuracy; note from the standard error statistics that accuracy differences are significant ($p < 0.05$).

C statistic: We investigated six C-models, where the first only uses C, the rest use C and the r'th uses $\{E_1 \dots E_r\}$. Table 3.2 shows the cross-validated accuracy of the models and whether they are admissible or not. There is no significant difference in accuracy between the two admissible models; we choose the larger model $r = 1$.

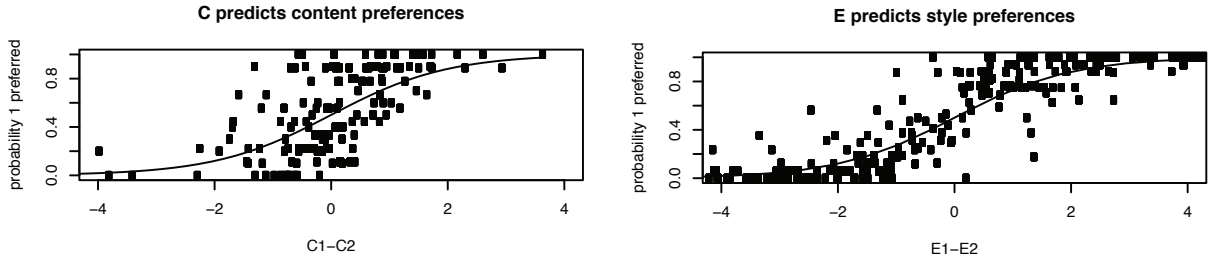


Figure 3.2: *Both E and C statistics are calibrated to user preferences in a comparison. Users will prefer image 1's style (resp. content) over image 2's style (resp. content) with a probability that depends on the difference in E (resp, C) values. On the* **left***, the predicted probability of preferring image 1 as a function of C1-C2 in a content experiment. Scattered points are true user observations; for most pairs, we have 9 or more observations. On the* **right***, the predicted probability of preferring image 1 as a function of E1-E2 in a style experiment. Scattered points are true user observations; for most pairs, we have 16 or more observations. Accuracy is confirmed by cross-validated estimates of classification accuracy (Sec 3.3).*

3.4 USER STUDY DETAILS

We conduct with two rounds user studies. The first round was a pilot study to produce usable data. The second round produced more data on style preferences. The first round had 300 image pairs for E-test and 150 image pairs for C-test, each of which was generated using Gatys method. For the E-test we randomly selected two transfered images from the same style and the same content but with different optimization parameters, then paired and displayed them in random order. For the C-test we follow the same process and only used pairs where the E statistic was in the top quartile.

For each task, users are presented with a question, an original image (style image for E-test and content image for C-test) and a transferred pair. Users are asked to choose a preferred image based on the displayed question. Overall, 16 users finished E-test, and 9 finished C-test task. From the first round we obtained 4854 clicks for E-test and 1410 clicks for C-test.

In the second round, to calibrate E regardless of transfer methods, we used a mixture of 939 image pairs generated from Universal (352), XL (294) and Gatys (294) methods (see methods explanation in Sec. 3.5.1). The style and content images are reused but the same style-content combination is not repeated. In total 24 users (a few also participated the first round) participate the second round and contributed 2232 clicks.

3.5 COMPARING STYLE TRANSFER METHODS WITH E AND C

With calibrated, meaningful measures of effectiveness and coherence, we can evaluate style transfer algorithms. We consider which algorithm is "best" and the effect choice of style has on performance. For analyzing the effects of weights, choice of style,and optimization objectives etc. we use the following procedure. We regress E (resp. C) for many style transfers produced by the algorithm of interest, then extract information from the coefficient weights.

3.5.1 Details

Style transfer methods compared:

Gatys ([1] and described above); we use the implementation by Gatys [1].

Gatys aggressive ([1] and described above); we use the same Gatys implementation, but with the aggressive weighting set.

Gatys, with histogram loss: as advocated by [4], we attach a histogram loss to Gatys method.

Gatys, with layerwise style weights: the style weight is varied by layer; we multiple style losses of layers by factors 64^{-2},128^{-2},256^{-2},512^{-2},512^{-2} respectively.

Gatys, with mean control: Gatys' loss, with an added L2 loss requiring that means in each transfer layer match to means in each style layer.

Gatys, with covariance control: replacing Gatys' gram matrix by covariant matrix.

Gatys, with mean and covariance control: replacing Gatys' style loss with losses requiring that means and covariances in each layer match.

Cross-layer: We used a *pairwise descending* strategy with pre-trained VGG-16 model. We use R11, R21, R31, R41, and R51 for style loss, and R42 for the content loss for style transfer.

Cross-layer, aggressive: as for XL, but with the aggressive weighting set.

Cross-layer, multiplicative (XM): A natural alternative to combine style and content losses is to multiply them; we form $L^m(I_n) = L_c(I_n, I_c) * L_s(I_n, I_s)$. This provides a dynamical weighting between content loss and style loss during optimization. Although this loss function may seem odd, it performs extremely well in practice.

Cross-layer, with control of covariance (XLC) Cross-layer loss, but replacing cross-layer gram matrices by cross-layer covariance matrices.

Cross-layer, with control of mean and covariance (XLCM) XLC, but with an added loss requiring that means in each layer match.

[1] https://github.com/leongatys/PytorchNeuralStyleTransfer

Gatys, augmented Lagrangian method (GAL): We use the Gatys' loss, but rather than only using LBFGS to optimize, we decouple layers to produce a constrained optimization problem and use the augmented Lagrangian method to solve this (after the procedure in [29] for decomposing MRF problems). As XM, this works effectively as dynamical weighting and performs extremely well. Details in Appendix.
Universal Style Transfer (Universal):(from [17], and its Pytorch implementation [2].
Style control: the style image is resized to content size.
Content control: the content image.
Comparison data: We have built two datasets on a wide range of styles and contents, using 50 style images (see Appendix B) and the 200 content images from the BSDS500 test set. The *main set* is used for most experiments, and was obtained by: take 20 evenly spaced weight values in the range 50-2000; then, for each weight value, choose 15 style/content pairs uniformly and at random. The *aggressive weighting set* is used to investigate the effect of extreme weights. This was built by taking 20 weight values sampled uniformly and at random between 2000-10000; then, for each weight value, choose 15 style/content pairs uniformly and at random. For each method, we then produced 300 style transfer images using each weight-style-content triplet. For Universal [17], since the maximum weight is one, we linearly map *main set* weights to the zero-one range. Our samples are sufficient to produce clear differences in standard error bars and evaluate different methods.

3.5.2 Results

We compare methods by constructing style transfers for each element of our dataset (a tuple of `style`, `content`, and `weight`). We then visualize the E and C statistics on a plot. We show the mean and covariance ellipse for E and C for various methods in Fig. 3.3, 3.4 and 3.5. Generally methods with strong C have weak E and vice versa, and we expect a trade-off (this is a Pareto frontier). But for some methods, which are **inadmissible**, the mean E and the mean C are both weaker than those available with another method. For such method, one could obtain better performance in both E and C by passing to some other method. Note that this criterion is weak, because it looks at mean E and mean C, and the covariance might argue for using a method with inadmissible means. Comparing Fig. 3.3, 3.4 and 3.5 suggests that there is nothing to be gained by using an inadmissible method. Notice, in particular, that inadmissible methods tend to have large variance in C; one might get a good C, but one might also get a bad one.

[2] https://github.com/sunshineatnoon/PytorchWCT

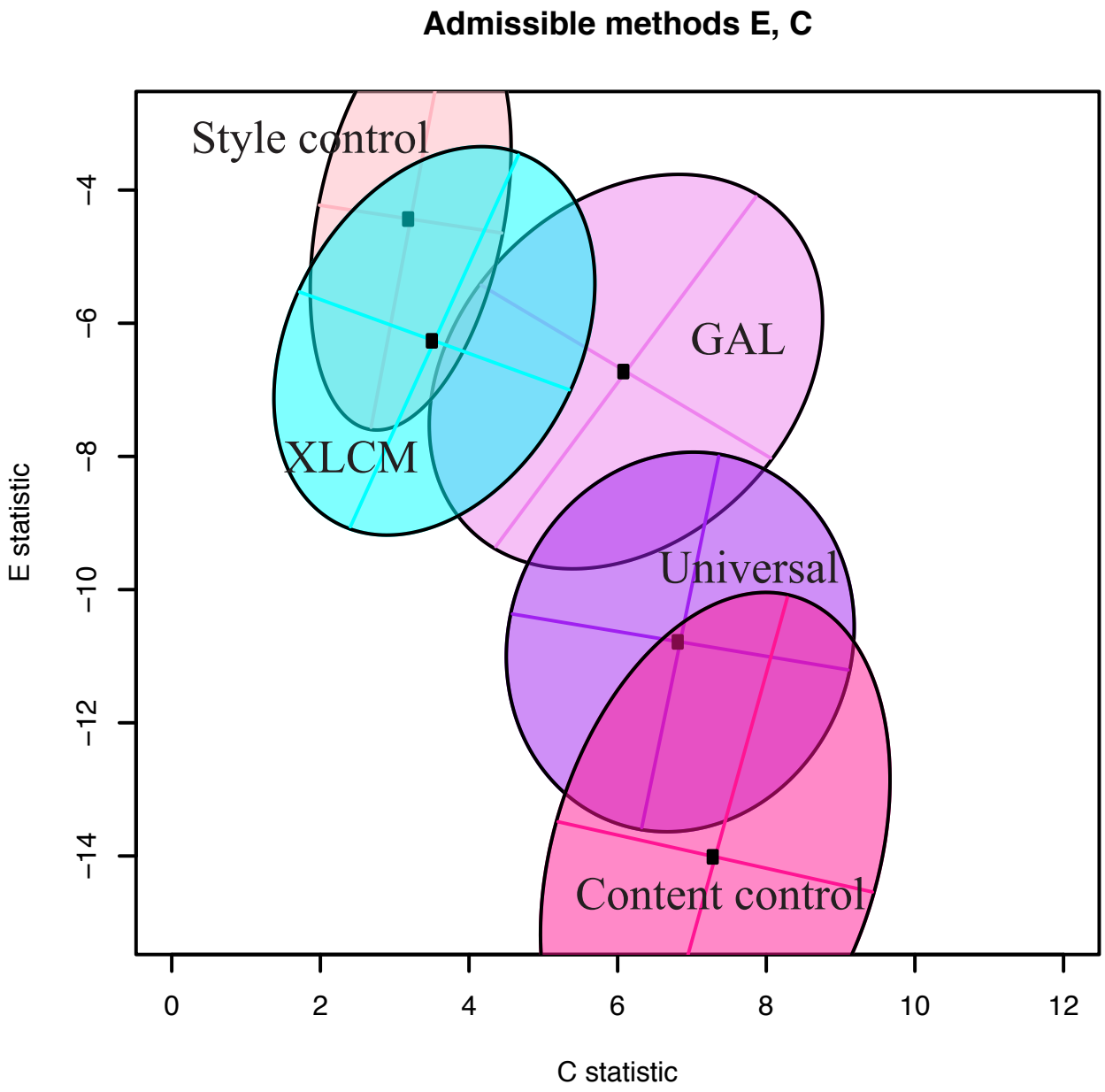


Figure 3.3: *E and C statistics for admissible methods. The plot shows mean (filled black circle) and 66% confidence ellipse, showing covariance of E and C values for each method. Notice: E and C are positively correlated, suggesting some dependence on either style (compare Fig. 3.6) or optimization difficulties; XLCM and GAL achieve better E, and universal achieves better C; controls are where expected (style control gets excellent E, weak C; content control weak E, excellent C).*

Method	Style Weight Effect	Significance (P-value)
XLCM	-0.40 (0.23)	0.05
GAL	-0.34 (0.19)	0.09
Universal	1.54 (0.89)	$< 1e-3$

Table 3.3: *We show the effect of style weight on E for admissible methods by multiplying the regression coefficient by the mean style weight (brackets show regression coefficient $\times$ standard deviation). This gives the range of differences in E caused by style weights. Note P-values are high for XLCM and GAL, so there is little evidence weights actually matter.*

Admissible methods: The style control has excellent E and weak C; the content control has excellent C and weak E. Each is admissible, because if one really wanted very strong E (resp. C) at all costs, one would use the style (resp. content) control. Universal style transfer has excellent C, but very weak E (i.e. the style is not much transferred, so the original image is quite coherent). Fig. 3.3 summarizes our data. Cross-layer methods do well. XLCM and GAL obtain only very slightly different E's, but different C's; although each is admissible, GAL should likely be preferred as it obtains a strong C with little erosion of E. U produces a better C, but at the cost of a markedly worse E. The differences between methods quite obviously achieve statistical significance (n=300; ellipses show covariance rather than standard deviation).

Inadmissible methods of the Gatys type are shown in figure 3.4, and of the cross-layer type are shown in figure 3.5. Note that XM is very close to being admissible.

Style and Weight: Style weights have surprisingly small effect on the E statistic (table 3.3). The choice of style is very important. Fig. 3.6 shows the result of regressing the E statistic against style identity; many styles are strongly advantageous or disadvantageous for many methods. There is no clearly dominant method here. It is obvious from the figure that any given method can be significantly advantaged by choosing the styles for transfer carefully. This is a trap for evaluators.

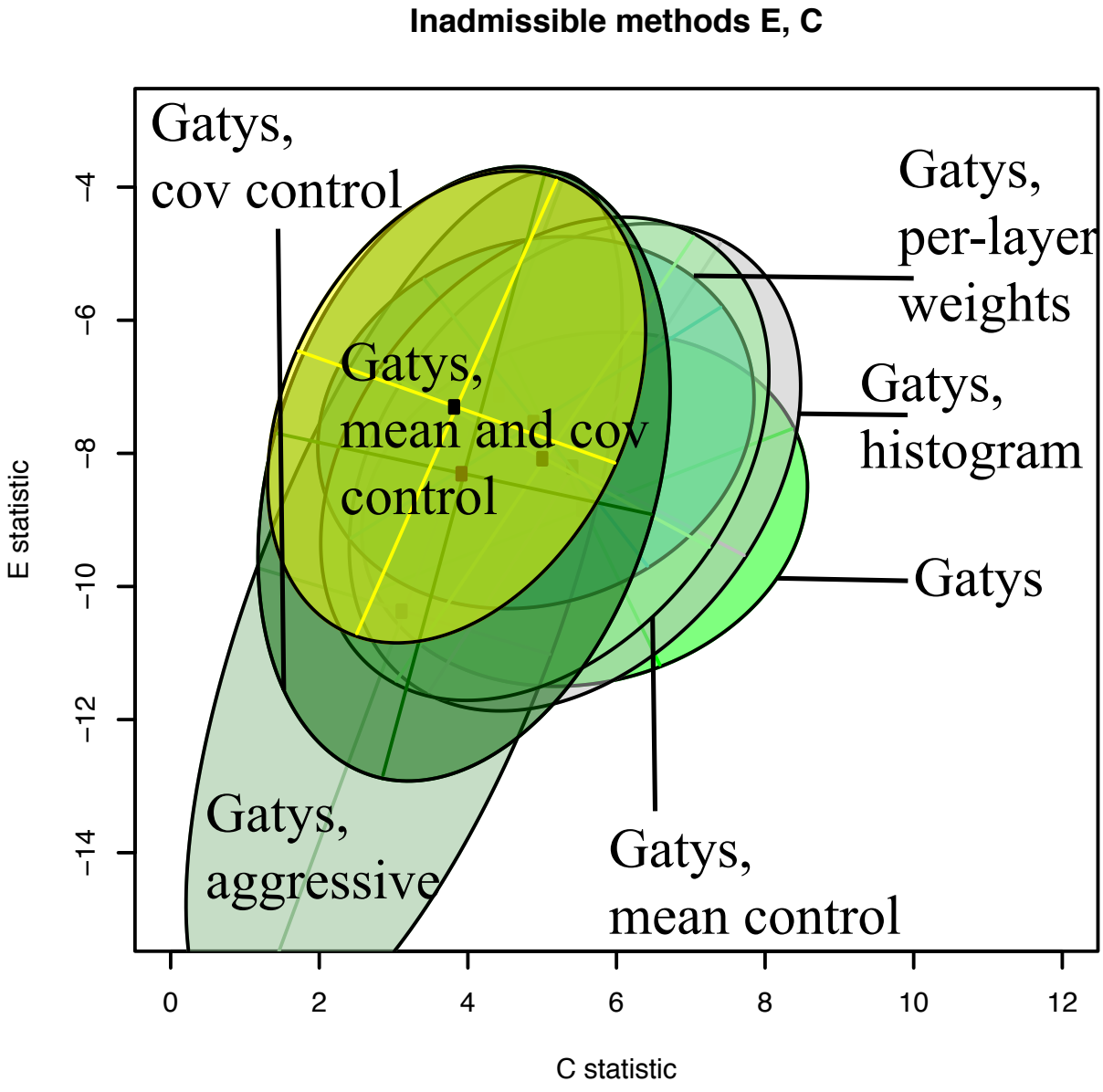


Figure 3.4: *E and C statistics for inadmissible methods of the Gatys type. The plot shows mean (filled black circle) and 66% confidence ellipse. Notice: E and C are positively correlated, suggesting some dependence on either style (compare Fig. 3.6) or optimization difficulties; the likely instability in Gatys' method is reflected by very high variance when an aggressive weight schedule is used.*

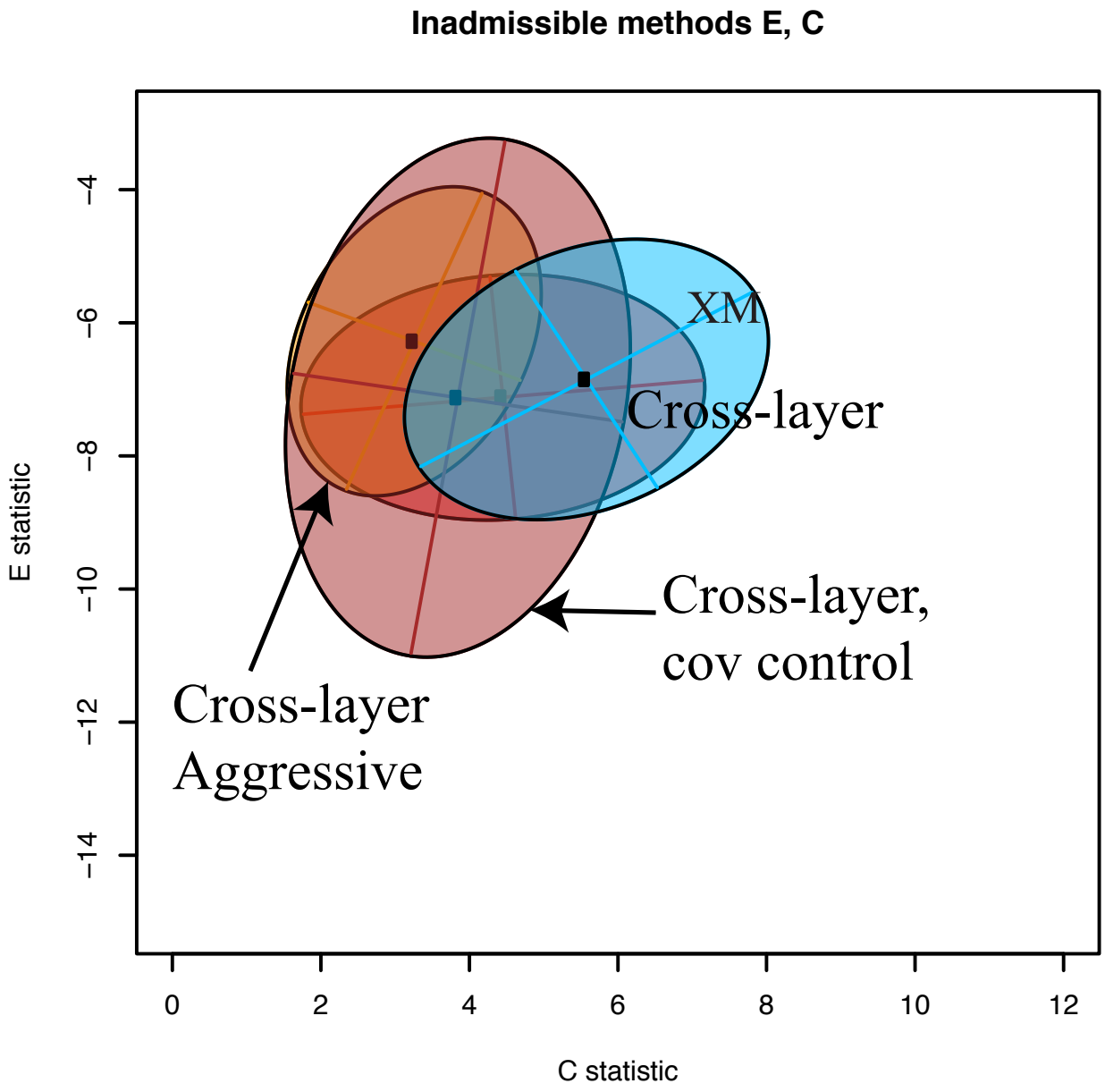


Figure 3.5: *E and C statistics for inadmissible methods of the cross-layer type. The plot shows mean (filled black circle) and 66% confidence ellipse. Notice: E and C are positively correlated, suggesting some dependence on either style (compare Fig 3.6) or optimization difficulties; the cross-layer method reacts to aggressive style weighting by producing increased E and lower C, as one would expect. XM performs best, and is very close to being admissible.*

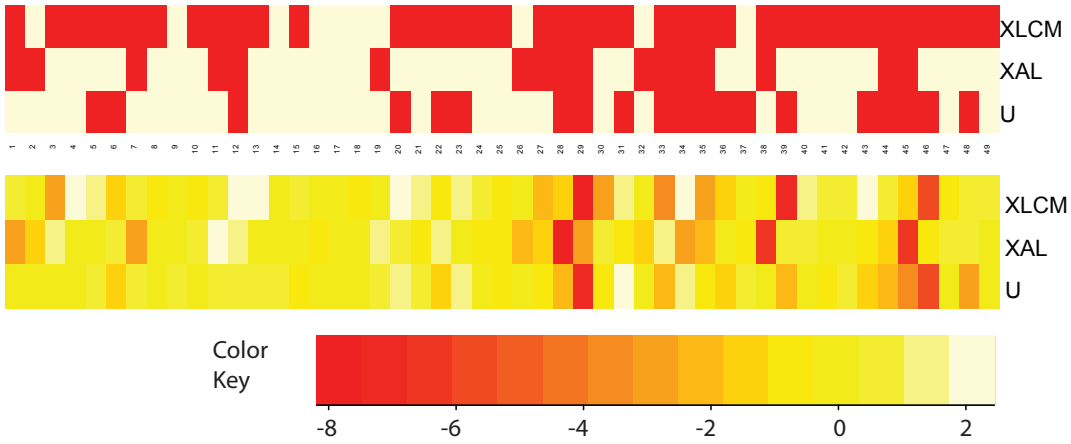


Figure 3.6: *The E measure that a method produces depends very strongly on the style; some styles transfer well, others poorly, even for admissible methods (this figure). On the* **top**, *a heatmap showing the significance of the dependency of the E statistic on style, red boxes indicate* $p < 0.05$ *(i.e. likely not an accident). Vertical coordinate gives the method, horizontal coordinate gives the style. While more detailed analysis would be required to reliably identify which styles have a strong effect of the method, it is clear that all methods are strongly affected by many styles. On the* **bottom**, *a heatmap showing the weight (positive=yellow means improves E; negative=red means weakens E) for each of our 50 styles for each method. All methods find some styles hard and others helpful.*

CHAPTER 4: DISCUSSION AND CONCLUSION

4.1 DISCUSSION

What causes the difference between Gatys' method and cross-layer losses? A **symmetry analysis** helps explain some aspects of our results. The Appendix D give a construction for all affine maps that fix the gram matrix for a layer and its parent (deeper networks follow the same lines). It is necessary to assume the map from layer to layer is linear. This is not as restrictive as it may seem; the analysis yields a local construction about any generic operating point of the network. In summary, we have: The between-layer gram matrix loss has very different symmetries to Gatys' (within-layer) method. In particular, the symmetry of Gatys' method rescales features while shifting the mean (because in this case $\mathcal{A}$ can contain strong rescaling with the right choice of $\mathbf{b}$). For the cross-layer loss, the symmetry cannot rescale, and cannot shift the mean. This implies that, if one constructs numerous style transfers with the same style using Gatys' method, the variance of the layer features should be much greater than that observed for the cross layer method. Furthermore, these symmetries impede optimization by making it hard to identify progress as massive changes in the input image may lead to no change in loss.

Increasing style weights in Gatys method should result in poor style transfers, by exaggerating the effects of the symmetry, and we observe this effect. Our construction casts light on part Gupta *et al.* 's observation linking large trace to instability. A small trace in the gram matrix implies many small eigenvalues. In turn, rescaling directions with small eigenvalues will change little unless very large scales are applied; but these correspond to very large shifts in the mean, which are difficult to obtain with current random start methods. However, a large trace in the gram matrix implies that there are many directions where a small shift in the mean will result in a small – but visible, because the eigenvalue is big – rescale from $\mathcal{A}$ that will lead to real changes, and so there is greater instability.

Our experimental evidence suggests the symmetries manifest themselves in practice. Gatys-like methods displays significantly larger variance in C than cross-layer methods, and aggressive weighting makes the situation worse. This suggests that the variance implied by the larger symmetry group is actually appearing. In particular, Gatys' symmetry group allows rescaling of features and shifting of their mean, which will cause the feature distribution of the transferred image to move away from the feature distribution of the style, causing the lower E statistic. Histogram regularization does not appear to help significantly.

Symmetries appear to interact strongly with optimization difficulties. GAL uses a stan-

dard optimization trick (insert variables and constraints to decouple terms in an unconstrained problem in the hope of making better progress with each step) and benefits significantly. In particular, GAL is largely immune to change in style weight (the coefficient is not significantly different to zero). This suggests that the main difficulty might lie with optimization procedures, rather than with losses.

4.2 CONCLUSION

We both qualitatively and quantatively show that the cross-layer gram-matrix has better performance than within-layer gram matrix. Style transfer methods have proliferated in the absence of a quantitative evaluation method. Our evaluation procedure attempts to provide evidents for strong style transfer methods. We calibrate out measurement to predict human preferences in style (resp. content) experiments, allowing extensive comparison of methods. Small variants on method – for example, changes to optimization procedure – seem to have significant effect on performance. This is a situation where quantitative evaluation is essential. Furthermore, our results suggest that the choice of style strongly affects the performance of all admissible algorithms.

APPENDIX A: SOME STYLE ALGORITHM DETAILS

A.1 QUICK OVERVIEW

Notice that in Fig 5 all Gatys related methods except *Gatys with mean and covariance control* have quite low E compared to the E for cross-layer methods in Fig 6. But *Gatys with mean and covariance control* has different symmetries to Gatys (because one is controlling both mean and covariance, rather than just the Gram matrix; the symmetries are like those of the cross-layer method). This suggests it is likely that the symmetry is at least part of the reason why some methods outperform others.

There are two possible reasons. First, the symmetry results in poor solutions being easy to find. Second, the symmetry causes optimization problems. Both issues appear to be in play. Figures 5 and 6 together suggest that methods have considerable variance in performance, which is consistent with poor solutions being easy to find. But the good performance of GAL (see Fig. 4) suggests that optimization is an issue, too.

Symmetries can create problems for optimization methods, because symmetries must be associated with strong gradient curvature at least some points. GAL uses a standard optimization trick to simplify the optimization problem; the success of this trick suggests that optimization of Gatys' loss is hard.

A.2 GAL

Gatys' loss is a function of feature values at each layer. One usually assumes that the feature values taken at layer l are a known function of the feature values at layer $l-1$. Here the function is given by the appropriate convolutional layer, etc. However, we could "cut" the network between layers, then introduce a constraint requiring that variables on either side of the cut be equal. We solve this constrained problem using the augmented lagrangian method (see [4] for this strategy applied to MRFs).

Write $f^l_{k,p}$ for the response of the k'th channel at the p'th location in the l'th convolutional layer; drop subscripts as required, and write $f^l = \phi^l(f^{l-1}_{.,.})$ for the function mapping layer to layer. GAL cuts the layers only at R41. We have not tried other cuts. It would be interesting to see what happened with more cuts, but the optimization problem gets big quickly. We introduce dummy variables $V_{k,p}$, and the constraint $V = \phi^4(f^3_{.,.})$. Write λ for lagrange multipliers corresponding to the constraint, I for the image, and $\lambda^{(i)}$ for the i'th estimate of those lagrange multipliers, etc.

The augmented lagrangian is now

$$\begin{aligned}\mathcal{L}(I,V,\lambda) &= \textstyle\sum_{l\neq 4} w_l L^l_{style}(I, I_{style}) \\ &\quad + w_4 L^4_{style}(V, I_{style}) \\ &\quad + L_{content}(V, I_{content}) \\ &\quad + L_{aug}(I,V,\lambda)\end{aligned} \tag{A.1}$$

where w_l is the style weight of each layer, L^l_{style} is the style loss for layer l, and $L_{content}$ is the content loss at R41, and

$$\begin{aligned}L_{aug}(I,V,\lambda) &= \tfrac{1}{KP}\textstyle\sum_{k,p}\Big(\lambda_l * (V_l - \phi^4(f^3_{:,:}(I))) \\ &\quad + \rho(V_l - \phi^4(f^3_{:,:}(I)))^2\Big)\end{aligned} \tag{A.2}$$

In the primal step, we first optimize the lagrangian with respect to I, using fixed V, λ using LBFGS. We then fix I, and optimize with respect to V (notice this involves solving a relatively straightforward linear system). The dual step then re-estimates the lagrange multipliers as usual:

$$\lambda_4^{(i+1)} = \lambda_4^{(i)} + \rho^{(i)}(V_4^{(i)} - f^4(I_n^{(i)})). \tag{A.3}$$

Finally, we update ρ by $\rho^{(i+1)} = 1.4\rho^{(i)}$.

A.3 CROSS-LAYER WITH CONTROL OF MEAN AND COVARIANCE (XLCM)

We observe that feature mean difference between I_s and I_c is directly related to the optimization performance of style transfer, e.g. when the content image have similar feature mean as style image the transfer image has better style quality. Therefore we introduce the L2 loss between each feature channel's mean of I_n and each feature channel's mean of I_s to enforce the transfer image has close feature mean to style image. Here is the loss for mean control.

$$L_{mean} = \sum_k \left(\sum_p \frac{f^l(I_n)}{P} - \sum_p \frac{f^l(I_s)}{P}\right)^2 \tag{A.4}$$

On the other hand, the covariant control is to replace cross-layer gram matrix by corre-

sponding cross-layer gram matrix with each feature subtracted by by its mean. Here is the new cross-layer loss with covariant control.

$$Cov_{ij}^{l,m}(I) = \sum_{p} \left[f_{i,p}^{l}(I) - \bar{f}_{i,p}^{l}(I) \right] \left[\uparrow f_{j,p}^{m}(I) - \uparrow \bar{f}_{j,p}^{m}(I) \right]^{T}. \tag{A.5}$$

Here $\bar{f}_{i,p}^{l}(I)$ is the tensor duplicated in p dimension with the mean of $f_{i,p}^{l}(I)$ over p.

APPENDIX B: SELECTED 50 STYLES

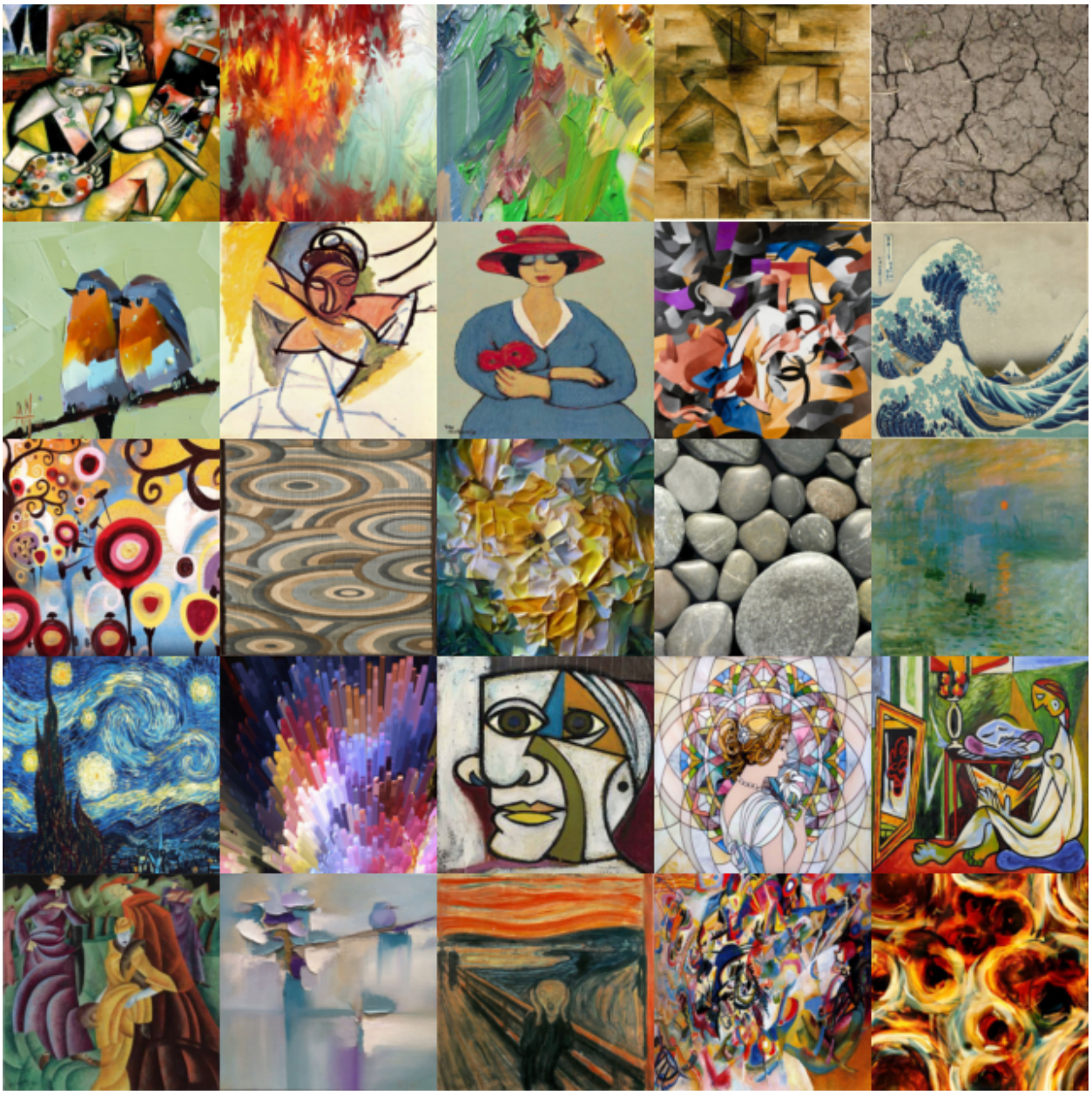


Figure B.1: *The first group of 50 styles.*

Figure B.1 and Figure B.2 display our 50 style images. Except the Universal style transfer, all other methods synthesize image from Gaussian noise with LBFGS optimizer. The content images and style images are resized to same width of 512 as the input for style transfers.

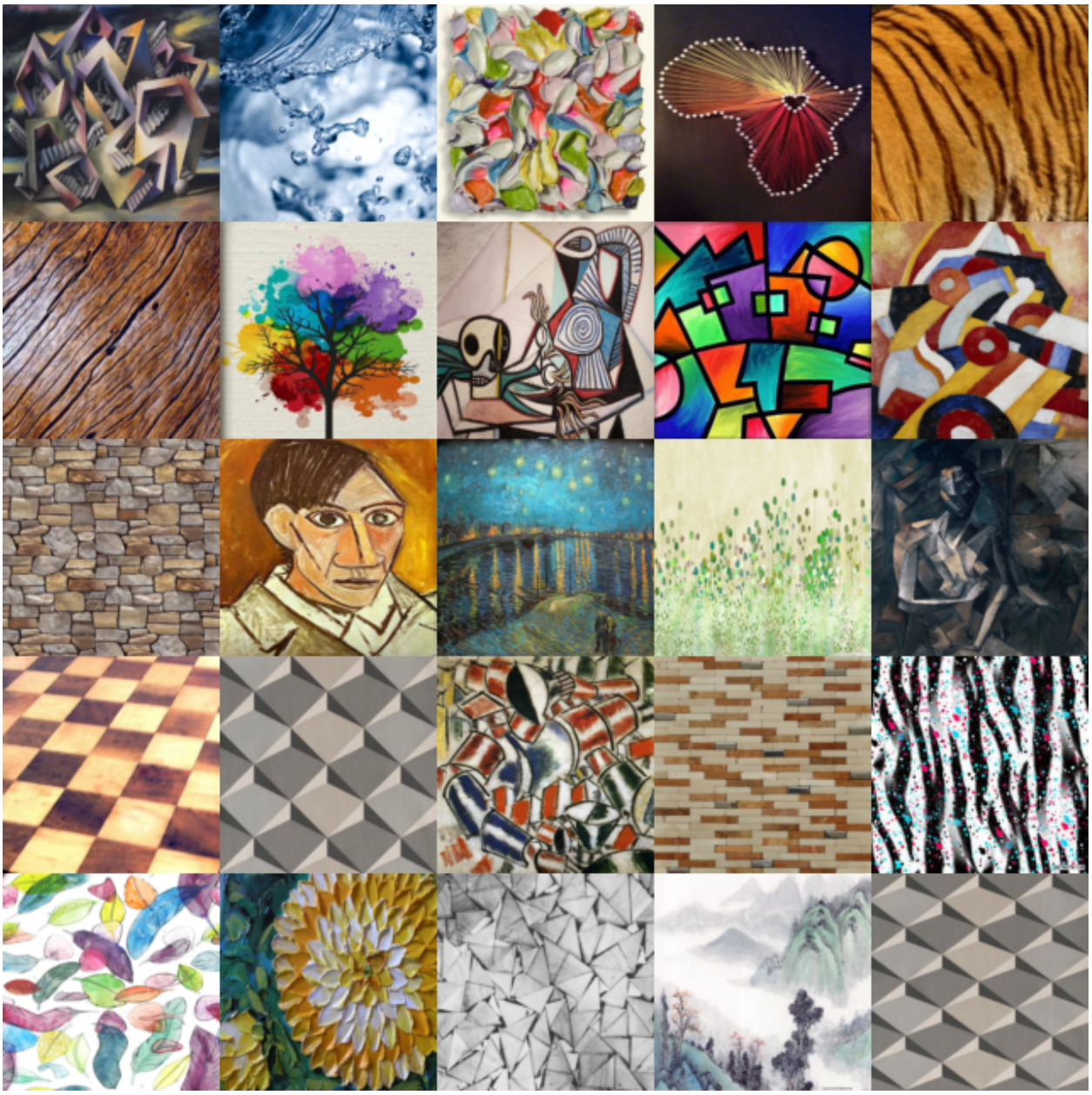


Figure B.2: *The second group of 50 styles.*

APPENDIX C: QUANTIZATION OF TRANSFERRED IMAGES UNDER USER STUDY REGRESSION MODELS

Recall in Section 4 of original text we regress base E and C statistic to user preference. We obtain one best E-model from E-test user preference, and one best C-model from that of C-test. These two models assign E and C scores for each transferred image (Sec. 4.1 of original text). Thus, we gather a scatter plot of all transferred images, and we quantize this scatter plot into a 3-by-3 grid, each cell has roughly same number of images. From this grid we generate a visualization of EC space (Fig.1 in original text).

This quantization shows similar trends with Figure 4-6 in the original text. Table C.1 shows the Top 5 methods ranking for all quantiles. In quantile of high C-score, high E-score, GAL is the top method. XM dominates both (middle C, middle E) and (high C, middle E), and Universal dominates both (middle C, low E) and (high C, low E). Other high E quantiles are dominated by cross-layer related methods. The worst quantile(low C-score,Low E-score) has Gatys aggressive as the most popular.

- GatysH – Gatys, with histogram loss
- GatysL – Gatys, with layerwise style weights
- GatysM – Gatys, with mean control
- GatysC – Gatys, with covariance control
- GatysCM – Gatys, with mean and covariance control
- XL – Cross-layer
- XM – Cross-layer, multiplicative
- XLC – Cross-layer, with control of covariance
- XLCM – Cross-layer, with control of mean and covariance
- GAL – Gatys, augmented Lagrangian method
- Universal – Universal Style Transfer

(low C-score, high E-score) **Cross-layer,aggressive:24.06%**, XLCM:20.92%, XLC:11.92%, XL:11.30%, GatysCM:9.21%	(middle C-score, high E-score) **XLC:14.56%**, Cross-layer,aggressive:13.60%, XLCM:13.41%, XL:13.22%, GAL:10.15%	(high C-score, high E-score) **GAL:25.56%**, XM:15.04%, XL:10.53%, GatysL:8.52%, GatysCM:6.77%
(low C-score, middle E-score) **GatysCM:15.29%**, GatysC:12.86%, Cross-layer, aggressive:11.65%, GatysL:11.65%, XLCM:8.50%	(middle C-score, middle E-score) **XM:11.69%**, GatysM:11.49%, GatysL:10.69%, GatysH:10.08%, GatysC:8.87%	(high C-score, middle E-score) **XM:15.45%,** GatysH:14.02%, Gatys:13.41%, GAL:13.01%, GatysM:11.18%
(low C-score, low E-score) **Gatys aggressive:23.97%**, GatysC:12.57%, XLC:10.02%, GatysCM:8.84%, GatysM:7.47%	(middle C-score, low E-score) **Universal:12.83%**, GatysH:10.73%, Gatys aggressive:10.47%, GatysM:10.21%, Gatys:9.69%	(high C-score, low E-score) **Universal:45.28%**, Gatys:15.75%, GatysH:7.87%, GatysM:6.69%, GatysL:4.53%

Table C.1: Top 5 methods ranking for each quantile under regression scores coordinate generated by selected E-model and C-model. Each transferred image has five E-statistic and one C-statistic, they are used to regress user preference in E-test and C-test (Sec. 4.1 in original text). Selected E and C models regress scores (higher is better) for each transferred image. We divide the scatter into 3-by-3 quantiles, and show method distribution for each quantile.

APPENDIX D: CONSTRUCTION OF AFFINE MAPS FOR SYMMETRY GROUPS

This difference in symmetry groups is important. Risser argues that the symmetries of gram matrices in Gatys' method could lead to unstable reconstructions; they control this effect using feature histograms. What causes the effect is that the symmetry rescales features while shifting the mean. For the cross-layer loss, the symmetry cannot rescale, and cannot shift the mean. In turn, the instability identified in that paper does not apply to the cross-layer gram matrix and our results could not be improved by adopting a histogram loss.

Write $\mathbf{x}_i$, (resp $\mathbf{y}_i$ for the feature vector at the i'th location (of N in total) in the first (resp second) layer. Write $\mathcal{X}^T = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, etc.

Symmetries of the first layer: Now assume that the first layer has been normalized to zero mean and unit covariance. There is no loss of generality, because the whitening transform can be written into the expression for the group. Write $\mathcal{G}(\mathcal{W}) = (1/N)\mathcal{W}^T\mathcal{W}$ for the operator that forms the within layer gram matrix. We have $\mathcal{G}(\mathcal{X}) = \mathcal{I}$. Now consider an affine action on layer 1, mapping $\mathcal{X}_1$ to $\mathcal{X}_1^* = \mathcal{X}_1\mathcal{A} + \mathbf{1}\mathbf{b}^T$; then for this to be a symmetry, we must have $G(\mathcal{X}_1^*) = \mathcal{A}\mathcal{A}^T + \mathbf{b}\mathbf{b}^T = \mathcal{I}$. In turn, the symmetry group can be constructed by: choose $\mathbf{b}$ which does not have unit length; factor $N(\mathcal{I} - \mathbf{b}\mathbf{b}^T)$ to obtain $\mathcal{A}(\mathbf{b})$ (for example, by using a cholesky transformation); then any element of the group is a pair $(\mathbf{b}, \mathcal{A}(\mathbf{b})\mathcal{U})$ where $\mathcal{U}$ is orthonormal. Note that factoring will fail for $\mathbf{b}$ a unit vector, whence the restriction.

The second layer: We will assume that the map between layers of features is linear. This assumption is not true in practice, but major differences between symmetries observed under these conditions likely result in differences when the map is linear. We can analyze for two cases: first, all units in the map observe only one input feature vector (i.e. 1x1 convolutions; the *point sample* case); second, spatial homogeneity in the layers.

The point sample case: Assume that every unit in the map observes only one input feature from the previous layer (1x1 convolutions). We have $\mathcal{Y} = \mathcal{X}\mathcal{M} + \mathbf{1}\mathbf{n}^T$, because the map between layers is linear. Now consider the effect on the second layer. We have $\mathcal{G}(\mathcal{Y}) = \mathcal{M}\mathcal{M}^T + \mathbf{n}\mathbf{n}^T$. Choose some symmetry group element for the first layer, $(\mathbf{b}, \mathcal{A})$. The gram matrix for the second layer becomes $\mathcal{G}(\mathcal{Y}^*)$, where $\mathcal{Y}^* = (\mathcal{X}\mathcal{A} + \mathbf{1}\mathbf{b}^T)\mathcal{M}^T + \mathbf{1}\mathbf{n}^T$. Recalling that $\mathcal{A}\mathcal{A}^T + \mathbf{b}\mathbf{b}^T = \mathcal{I}$ and $\mathcal{X}^T\mathbf{1} = 0$, we have

$$\mathcal{G}(\mathcal{Y}^*) = \mathcal{M}\mathcal{M}^T + \mathbf{n}\mathbf{n}^T + \mathbf{n}\mathbf{b}^T\mathcal{M}^T + \mathcal{M}\mathbf{b}\mathbf{n}^T \tag{D.1}$$

so that $\mathcal{G}(\mathcal{X}_2^*) = \mathcal{G}(\mathcal{X}_2)$ if $\mathcal{M}\mathbf{b} = 0$. This is relatively easy to achieve with $\mathbf{b} \neq 0$.

Spatial homogeneity: Now assume the map between layers has convolutions with max-

imum support $r \times r$. Write u for an index that runs over the whole feature map, and $\psi(\mathbf{x}_u)$ for a stacking operator that scans the convolutional support in fixed order and stacks the resulting features. For example, given a 3x3 convolution and indexing in 2D, we might have

$$\psi(\mathbf{x}_{22}) = \begin{pmatrix} \mathbf{x}_{11} \\ \mathbf{x}_{12} \\ \dots \\ \mathbf{x}_{33} \end{pmatrix} \tag{D.2}$$

In this case, there is some $\mathcal{M}$, $\mathbf{n}$ so that $\mathbf{y}_u = \mathcal{M}\psi(\mathbf{x}_u) + \mathbf{n}$. We ignore the effects of edges to simplify notation (though this argument may go through if edges are taken into account). Then there is some $\mathcal{M}$, $\mathbf{n}$ so we can write

$$\mathcal{G}(\mathcal{Y}) = (1/N)\sum_u \mathcal{M}\psi(\mathbf{x}_u)\psi(\mathbf{x}_u)^T\mathcal{M}^T + \mathbf{n}\mathbf{n}^T \tag{D.3}$$

Now assume further that layer 1 has the following (quite restrictive) spatial homogeneity property: for pairs of feature vectors within the layer $\mathbf{x}_{i,j}$, $\mathbf{x}_{i+\delta,j+\delta}$ with $|\,\delta\,| \leq r$ (ie within a convolution window of one another), we have $\mathbf{x}_{i,j}\mathbf{x}_{i+\delta,j+\delta} = \mathcal{I}$. This assumption is consistent with image autocorrelation functions (which fall off fairly slowly), but is still strong. Write ϕ for an operator that stacks $r \times r$ copies of its argument as appropriate, so

$$\phi(\mathcal{I}) = \begin{pmatrix} \mathcal{I} & \dots & \mathcal{I} \\ \dots & \dots\dots & \\ \mathcal{I} & \dots & \mathcal{I} \end{pmatrix}. \tag{D.4}$$

Then $G(\mathcal{Y}) = \mathcal{M}\phi(\mathcal{I})\mathcal{M}^T + \mathbf{n}\mathbf{n}^T$. If there is some affine action on layer 1, we have $G(\mathcal{Y}^*) = \mathcal{M}\left(\psi(\mathcal{A})\phi(\mathcal{I})\psi(\mathcal{A}^T) + \psi(\mathbf{b})\psi(\mathbf{b}^T)\right)\mathcal{M}^T + \mathbf{n}\mathbf{n}^T$, where we have overloaded ψ in the natural way. Now if $\mathcal{M}\psi(\mathbf{b}) = 0$ and $\mathcal{A}\mathcal{A}^T + \mathbf{b}\mathbf{b}^T = \mathcal{I}$, $\mathcal{G}(\mathcal{Y}^*) = \mathcal{G}(\mathcal{Y})$.

The cross-layer gram matrix: Symmetries of the cross-layer gram matrix are very different. Write $\mathcal{G}(\mathcal{X}, \mathcal{Y}) = (1/N)\mathcal{X}^T\mathcal{Y}$ for the cross layer gram matrix.

Cross-layer, point sample case: Here (recalling $\mathcal{X}^T\mathbf{1} = 0$)we have $\mathcal{G}(\mathcal{X}, \mathcal{Y}) = \mathcal{M}^T$. Now choose some symmetry group element for the first layer, $(\mathcal{A}, \mathbf{b})$. The cross-layer gram matrix becomes

$$\begin{aligned} \mathcal{G}(\mathcal{X}^*, \mathcal{Y}^*) &= (1/N)(\mathcal{A}\mathcal{X}^T + \mathbf{b}\mathbf{1}^T)\left[(\mathcal{X}\mathcal{A}^T + \mathbf{1}\mathbf{b}^T)\mathcal{M}^T + \mathbf{1}\mathbf{n}^T\right] && \text{(D.5)} \\ &= \mathcal{M}^T + \mathbf{b}\mathbf{n}^T && \text{(D.6)} \end{aligned}$$

(recalling that $\mathcal{A}\mathcal{A}^T + \mathbf{b}\mathbf{b}^T = \mathcal{I}$ and $\mathcal{X}^T\mathbf{1} = 0$). But this means that the symmetry requires $\mathbf{b} = \mathbf{0}$; in turn, we must have $\mathcal{A}\mathcal{A}^T = \mathcal{I}$.

Cross-layer, homogeneous case: We have

$$\mathcal{G}(\mathcal{X}, \mathcal{Y}) = (1/N) \sum_u \mathbf{x}_u \left[\psi(\mathbf{x}_u)^T \mathcal{M}^T + \mathbf{n}^T\right] = \mathcal{M}^T. \tag{D.7}$$

Now choose some symmetry group element for the first layer, $(\mathcal{A}, \mathbf{b})$. The cross-layer gram matrix becomes

$$\begin{aligned} \mathcal{G}(\mathcal{X}^*, \mathcal{Y}^*) &= (1/N) \sum_u \Bigg\{ (\mathcal{A}\mathbf{x}_u + \mathbf{b}) \\ &+ \left[\left(\psi(\mathbf{x}_u)^T \psi(\mathcal{A}^T) + \psi(\mathbf{b})\right) \mathcal{M}^T + \mathbf{n}^T\right] \Bigg\} \\ &= \mathcal{M}^T + \mathbf{b}\mathbf{n}^T \end{aligned} \tag{D.8}$$

(recalling the spatial homogeneity assumption, that $\mathcal{A}\mathcal{A}^T + \mathbf{b}\mathbf{b}^T = \mathcal{I}$ and $\mathcal{X}_1^T\mathbf{1} = 0$). But this means that the symmetry requires $\mathbf{b} = \mathbf{0}$; in turn, we must have $\mathcal{A}\mathcal{A}^T = \mathcal{I}$.

REFERENCES

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[2] R. Novak and Y. Nikulin, "Improving the neural algorithm of artistic style," *arXiv preprint arXiv:1605.04603*, 2016.

[3] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2011.

[4] P. Wilmot, E. Risser, and C. Barnes, "Stable and controllable neural texture synthesis and style transfer using histogram losses," *arXiv preprint arXiv:1701.08893*, 2017.

[5] J. B. Tenenbaum and W. T. Freeman, "Separating Style and Content with Bilinear Models," *Neural Computation*, vol. 12, no. 6, pp. 1247–1283, 2000. [Online]. Available: http://www.mitpressjournals.org/doi/10.1162/089976600300015349

[6] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*, pp. 341–346, 2001. [Online]. Available: http://portal.acm.org/citation.cfm?doid=383259.383296

[7] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*, no. August, pp. 327–340, 2001. [Online]. Available: http://portal.acm.org/citation.cfm?doid=383259.383295

[8] J. D. Bonet, "Multiresolution sampling procedure for analysis and synthesis of texture images," *SIGGRAPH*, 1997.

[9] E. P. Simoncelli and J. Portilla, "Texture characterization via joint statistics of wavelet coefficient magnitudes," in *ICIP*, 1998.

[10] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 262–270. [Online]. Available: http://papers.nips.cc/paper/5633-texture-synthesis-using-convolutional-neural-networks.pdf

[11] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

[12] X. Wang, G. Oxholm, D. Zhang, and Y.-F. Wang, "Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer," *arXiv preprint arXiv:1612.01895*, 2016.

[13] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stylebank: An explicit representation for neural image style transfer," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[14] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *ICLR*, 2017. [Online]. Available: https://arxiv.org/abs/1610.07629

[15] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *CoRR*, vol. abs/1607.08022, 2016. [Online]. Available: http://arxiv.org/abs/1607.08022

[16] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," *arXiv preprint arXiv:1703.06868*, 2017.

[17] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," *arXiv preprint arXiv:1705.08086*.

[18] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang, and J. Kautz, "A closed-form solution to photo-realistic image stylization," *arXiv preprint arXiv:1802.06474*, 2018.

[19] T. Q. Chen and M. Schmidt, "Fast patch-based style transfer of arbitrary style," *arXiv preprint arXiv:1612.04337*, 2016.

[20] Y. Shih, S. Paris, C. Barnes, W. T. Freeman, and F. Durand, "Style transfer for headshot portraits," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 1–14, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2601097.2601137

[21] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep Photo Style Transfer," 2017. [Online]. Available: http://arxiv.org/abs/1703.07511

[22] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman, "Controlling perceptual factors in neural style transfer," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[23] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," *arXiv preprint arXiv:1701.01036*, 2017.

[24] A. J. Champandard, "Semantic style transfer and turning two-bit doodles into fine artworks," *arXiv preprint arXiv:1603.01768*, 2016.

[25] Y. Jing, Y. Yang, Z. Feng, J. Ye, and M. Song, "Neural style transfer: A review," *arXiv preprint arXiv:1705.04058*, 2017.

[26] A. Gupta, J. Johnson, A. Alahi, and L. Fei-Fei, "Characterizing and improving stability in neural style transfer," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4087–4096.

[27] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1, pp. 503–528, 1989.

[28] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3319–3327.

[29] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.